- What are the precision and accuracy of the measurements used in the study?
- Did the study actually measure what it claimed to?

The questions in Checklist D would focus on the fundamental questions:

- Has the data set captured the variability within the population of interest?
- Is it sufficient in size and quality to support the estimate, decisions, or actions recommended in this risk assessment?
- Can we quantify potential departures of our estimates from their correct (but unknown) values? Why and how?

Each of the bulleted items above has some detailed questions associated with it.

1.2 Tiered risk assessments

There is no subset of questions that can be selected since it cannot be foreseen which question is critical to evaluating a particular study. However, there is a basis for limiting the effort needed to establish representativeness. First, materially unimportant variables-as established, for example, by a sensitivity analysis-need not be fully addressed. Second, many of the checklist questions are relevant when variability and extreme percentiles must be characterized; they become less consequential when only a central tendency need be assessed. Finally, for a screening risk assessment, only qualitative degrees of representativeness are needed. For example, if it is known only that study results will conservatively overestimate exposures, then that study could be useful for a screening level risk assessment, but probably not for subsequent tiers.

2) Sensitivity

There are two kinds of sensitivity in a probabilistic calculation. They are related to the distinction between variability and uncertainty. We may, with some loss of generality,

suppose that the calculation is a determined procedure F that processes a collection S = {p1 , p2, . . . , pN} of "inputs," each of which is a (possibly degenerate) probability distribution, and outputs a single probability distribution F(S). If there is a material change in inferences based on F(S) when one of the input distributions, say pl, is collapsed to a point, then the calculation is sensitive to the variability in pl. Otherwise, the distribution pl can, with some safety, be replaced by a single number (a degenerate distribution).

Uncertainty in the input pl can often be described as a collection of possible distributions {pl'} that are "close" to pl in some sense. A typical example is when pl is parametric and {pl'} is described by a set of alternate values of the parameters. There may even be a probability distribution on {pi'} (a Bayesian "prior"). If, by replacing pl by an arbitrary element of {pl'}, the inferences based on F(S) change in a material way, then the calculation is sensitive to the uncertainty in pl.

The data must be sufficient to establish either that a variable is not a sensitive input or, if it is, the data must be sufficient to characterize the variability or the uncertainty or both, depending on which contribute to the sensitivity. This provides one basis for deciding when data are adequate. However, it could be argued that any data acceptable for use in a screening risk assessment are necessarily acceptable in subsequent tiers-at a cost.

To be specific, for data to be acceptable at all they must provide some valid information about the population of interest and some quantifiable level of uncertainty must be established (no matter how great that level is). This is true for any risk assessment at any tier, not just for probabilistic risk assessments. For screening use, inputs would have to be set at extreme (but realistic) levels consistent with the data and their uncertainty, in such a way as to ensure a "conservative" estimate of risk-that is, one biased high. Once this is accomplished, it would seem there is no obstacle to using the

same data in the same way in subsequent tiers, with the price for doing so being estimates that are still biased high.

3) Adjustments

Geostatistical methods are available for certain adjustments of spatial scales. Good references are Cressie, N. “Statistics for Spatial Data;” Joumel, A. and C. Huijbregts, “Mining Geostatistics.” In particular, methods such as “conservation of lognormality” have been developed to adjust for differences in spatial measurement scale (this has been termed the “change of support’ problem). This is the spatial analog of the DW model.

Adjustments should be applied with extreme caution because results can be very sensitive to them. Similarly, surrogate data should be used very cautiously. A good point of departure for considering adjustments is the following definition, constructed to capture the use of “representative” in EPA guidance (“Guiding Principles for Monte Carlo Analysis, EPA/630/R-97/001):

> Data are “representative” when they admit objective and quantifiable statements concerning the accuracy of the relevant inferences made from them.

From this point of view, adjustments can be considered (and defended) when made in a way that allows the potential bias or imprecision thereby introduced to be quantified in the risk assessment.

EDFs (Issue Paper #2)

1) *Selecting an EDF or PDF*

The primary consideration is the effect the choice will have on the risk assessment results. Each choice has relative advantages and disadvantages. They come down to this: using the EDF honors the data but subjects the calculation to the risk that the EDF poorly represents population variability and percentiles, a risk that can sometimes be decreased by using a well-chosen PDF. Using a PDF requires some theory and professional judgment and subjects the calculation to the risk that either (or both) could be wrong or inapplicable.

The choice is not inherently one of preference. With small data sets especially, an EDF is unlikely to represent an upper percentile adequately and so is manifestly a bad choice. (That's not to say that any particular PDF fit to the data is necessarily better!) When measurement error is large, the EDF will not appropriately separate variability and uncertainty. On the other hand, when the data set is large and not fit well by any theoretical distribution function, using the EDF is an excellent approach.

So we come back to the basic point: what effect will choice of distribution function(s) have on the risk assessment results? This is determined in part by sensitivity analysis. For this, the exponential tail fitting approach is particularly intriguing, because it seems to provide a robust opportunity to explore how relatively more or less extrapolation beyond the sample maximum (or minimum) will influence the results.

2) Goodness of Fit

The best basis for concluding that a fitted distribution adequately represents a data set is when (1) there is a theoretical reason to presuppose the data will be represented by such a distribution and (2) the fit is consistent with that presupposition. In this situation, P-values are meaningful and useful provided that one appropriate goodness-of-fit (GOF) test is chosen before obtaining and testing the data.

Graphical examination of the distribution is crucial. All empirical distributions will depart from the theoretical fit, so the nature and amount of departure must be assessed. It is highly unlikely that any standard GOF test will produce P-values that reflect the sensitivity of the risk assessment results to these departures. In particular, goodness of fit in the upper (sometimes lower) percentiles is usually far more important than goodness of fit elsewhere.

In many cases, where many input variables are involved in a risk calculation, using fitted distributions that reproduce the means and variances of the data is likely to produce adequate results. So, more than any P-value or selection of GOF test, these three criteria will be practically useful for risk assessments:

1. Correctly represent the centers (means and medians) of the input distributions,
2. Correctly represent the variances of the input distributions.
3. Fit the important tails of the data as well as possible.

(The “important tails” are the tails most influencing the upper percentile risk estimates. The definition of the tail-e.g., data beyond what percentile-will depend on which upper percentiles are being characterized in the risk assessment.) Note that EDFs will satisfy the third criterion only when data sets are large enough to estimate extreme percentiles with confidence.

When only summary statistics are available, there is an inherent problem in fitting any distribution: it is impossible to estimate uncertainty. Using additional information about possible limits to the data (that is, what the most extreme values could be), one should over-estimate the amount of uncertainty in the fit and use that in a sensitivity analysis. Uncertainty in the variance of the data is particularly important for probabilistic risk assessments.

When the better known distributions do not fit the data, there is exceptionally little advantage to resorting to someone's system of distributions, such as the generalized F. First, there is usually no theoretical basis for adopting any of these distributions. Second, there is little assurance that the best fitting distribution in a family will adequately represent what is of importance, namely the variance and tails. Third, reproducing the calculations can be difficult if the family of distributions is not in general use or is ad-hoc, like the five-parameter generalized F distribution is. Fourth, many of these families of distributions include obscure members whose estimation theory might not be well understood or even known. It would be better for the risk assessor to work with familiar constructs whose properties (especially with regard to influencing the risk assessment outcome) are well known.

3) Uncertainty

Every standard method of assessing uncertainty has limitations. Maximum likelihood methods often are based on asymptotic normality, which sometimes is not achieved even for impractically large data sets. There are applications where the bootstrap does not work-it is not theoretically justified. Certain methods, such as pretending the likelihood function is a probability distribution, simply have no justification (based on the theory of estimation).

In general, uncertainty should be assessed as aggressively as possible. As many possible contributors to uncertainty should be considered and as many of these as

possible should be incorporated in the risk assessment, because their effects accumulate.

An excellent method for assessing uncertainty is to randomly divide datasets into parts, perform calculations (such as fitting distributions, estimating statistics, and computing risk) based on each part, and evaluate the differences that arise. Certain forms of the bootstrap and its relatives, such as the jackknife, automate parts of this procedure.

Robert C. Lee, Golder Associates Inc.

Comments Regarding "Issue Paper on Evaluating Representativeness of Exposure Factors Data"

1. The issue of representativeness relates to how the risk assessor makes judgments and corrections regarding uncertainty inherent in a nonrepresentative sample. Discussion of the differences between uncertainty (bias and/or error) and variability (heterogeneity) would be useful to avoid confusion. For example, Checklist I misleadingly implies that measurement error can have an effect on variability, which is an inherent property of a population.

Uncertainty can either be characterized as systematic (bias) or nonsystematic (error). Uncertainty in exposure assessment may stem from:

Model errors

Errors in the design of the assessment method (i.e. measure of exposure)

Errors in the use of the method

Subject limitations

Analytical errors

One way to represent bias and error is as follows. A measured or observed value X_i can be represented as a function of the true value T_i, bias b, and nonsystematic error E_i, as:

$$X_i = T_i + E_i + b$$

The population distribution of *Ts* represents variability. However, perfect knowledge is rarely available. Therefore, E can be represented, for example, as a normal distribution with a mean of zero and variance as:

$$\sigma^2_E = \sigma^2_X - \sigma^2_T$$

where o^2_x is the variance of the uncertain measure X, and o^2_T is the true variance (assuming independence).

Bias (which can be positive or negative) can be represented as a deterministic shift in the mean of X as compared to the mean of *T*, as:

$$\mu_b = \mu_X - \mu_T$$

Thus, error and bias can have an effect on the estimated population distribution, but not on the true variability.

2. In many cases, an approach that uses "reference individuals" or strata rather than attempting to evaluate or estimate variability in a broad population may be useful. For instance, if one is concerned about children's exposure to lead in a Western mining town, it may be simpler as a first step to hypothesize a few examples of children with deterministic characteristics with regard to site-specific population variability, and then evaluate the uncertainty associated with these reference individuals exposures. This method can be relatively inexpensive and easy compared to population sampling, and could be used as a screening step in an iterative decision-making framework.

3. The exact meanings of the terms "probability sample" and "probability sampling" as used in the issue paper are unclear. Presumably these are broad terms covering schemes such as random, stratified, cluster, composite, etc. sampling. If so, then there should be clarification and discussion regarding the methodological and inferential differences between these methods. For example, simple random sampling may not be appropriate for all environmental exposure variables. If an exposure factor varies geographically, then it may be more appropriate to spatially stratify the population, and characterize the factor within each strata as accurately and precisely as possible.

4. **As** stated in the text (page 8, final paragraph), the process of determining the "importance of discrepancies and making adjustments" may be highly "subjective". However, the remainder of the discussion focuses heavily on frequentist methods of accounting for sources of uncertainty, which may not be the most appropriate approach. There should be discussion regarding both empirical and nonempirical Bayesian methods of population inference, since these methods are very powerful and are increasingly used in risk applications. A major advantage of Bayesian methods is that they allow refinement or "updating" of a priori knowledge with additional data or information.

5. More attention is devoted to "temporal" characteristics of a population than "individual" or "spatial" characteristics in the text. The reason for this is unclear. There should be discussion of how to determine the relative importance of these characteristics in risk assessment.

6. Discussion of Bayesian techniques may be useful in Section 5 of the paper, which covers issues involved with improving representativeness.

7. Discussion of the use of simulations for future scenarios would be useful, For example, if a the characteristics of a population are changing over time, time trends could be incorporated into a simulation to determine the parameters of an particular exposure variable in, say, 20 years.

Comments Regarding "Issue Paper on Empirical Distribution Functions and Nonparametric Simulation"

1. The assumptions listed in the Introduction of the Issue Paper are important and should be discussed further. The first assumption,“. . data are sufficiently representative of the exposure factor in question”, is rarely met. Uncertainty associated with representativeness is often considerable. The second assumption, "...the analysis involves and exposure/risk model which includes additional exposure factors”, is often true, although evaluation of the upper tail of a variability distribution is often difficult because of its uncertainty. If the tail is of interest, it may be preferable to stratify the analysis so that the mean of a high-exposure stratum can be used in the risk assessment. The third assumption, "... Monte Carlo methods will be used to investigate the variation in exposure/risk”, may be true in practice, but other simple analytical and numerical methods exist. Given simple distributional assumptions (e.g. lognormality), a hand calculator can be used to calculate probabilistic output of many regulatory risk assessment models.

2. Examples of EDFs that have been used in risk assessments would be useful.

3. The statement implying that it is rare that theoretical probability distribution functions are “available” for exposure factors deserves discussion. For example, under the maximum-entropy criterion, theoretical PDFs may be fit in a rigorous manner using various combinations of limited a *priori* information. Furthermore, the assumption of lognormality for many exposure variables and models has a theoretical as well as a mechanistic basis. It is hard to argue against using lognormal distributions when non-negative, unimodal, positively skewed data are available.

Regardless, there is a practical continuum between using an EDF and, say, a maximum-entropy theoretical distribution. The issue of sensitivity is important; i.e. when does it make a difference in a risk assessment? In general, EDFs may take more time to develop. Discussions of the utility of particular distributions should be separated from theoretical arguments. An iterative approach to refinement of environmental

exposure distribution functions should be discussed. This could potentially avoid inefficiency, and could be used to focus research dollars. If conducted within a Bayesian framework, prior EDFs or PDFs can be refined given additional data.

4. Much discussion in the text centers on the appropriateness of particular goodness-of-fit methods, visualization, etc. All of these methods are “blunt tools”. Most statisticians simply use a number of different methods simultaneously or iteratively. If all the methods agree that a particular parametric distribution "fits" the data, then that distribution is probably appropriate. If they disagree, then the mechanistic and statistical justification for a particular distribution form and the sensitivity of the model output to the distribution defined should be examined; an EDF may be more appropriate. If the model output is insensitive to the particular PDF defined for a particular variable, then it probably does not matter what shape it takes.

Samuel Morris

Comments on Issue Paper on Evaluating Representativeness of Exposure Factors Data

3.1 Inferences from a sample to a population

The population of concern at a Super-fund site is generally the population surrounding the site. This is true if the concern is for exposures during remediation activities. If there is some residual risk that may last over an extended time, the population of concern may change. In a brownfields situation, for example, the population of concern may be people who will work at the site years into the future. These people may be quite different than the population currently living around the site.

4. COMPONENTS OF REPRESENTATIVENESS

There is no question that one would like a clear definition of the population of concern, but if a representative sampling of the characteristics of that population has not been done, that definition doesn't exist. Isn't that why one uses information from a surrogate population? That question then is, if one cannot characterize the population of concern, how can one know if the surrogate population is suitable to represent the population of concern? The answer is a practical one. It depends on the availability of resources, which in turn one hopes depends on how severe the risk is judged to be.

4.1 Internal components - surrogate data versus the study population

Certainly the representativeness of the surrogate study for its own study population should be evaluated. This paragraph seems to suggest that every assessor that makes use of a surrogate study should make this evaluation. Good surrogate studies are generally used over and over again by many assessors. Such an evaluation should only need to be made once, with the results made available to all assessors. Along with this evaluation should be an evaluation of the character of the population for which

the particular surrogate study is useful. This could go further to provide some limiting population characteristics beyond which the surrogate would not be recommended.

4.2 External components - population of concern versus surrogate population

The suggestion of using several national Food Consumption Surveys as a basis to extrapolate dietary habits into the present or future seems like a rather precarious thing to do. It also is something that could only be done for an extremely large, important, and well-funded assessment. It is another study that, if done at all, should only be done once and results made available widely.

Regarding several assessors independently speculating on the mean and coefficient of variation of a parameter (expert judgment?), to avoid the phenomenon of anchoring, a useful protocol is to have the experts begin from the extremes and probabilities toward the central point, rather than beginning with the mean.

Checklist I.

I don't understand the questions, "For what population or subpopulation size was the sample size adequate for estimating measures of central tendency . . .and other types of parameters?" The previous questions ask if the sample size was adequate, etc. Presumably this means it is adequate for the size of the population that was studied. I am assuming that this checklist pertains to an internal analysis of the surrogate study and has nothing at this point to do with a different population that is of concern to the assessor.

Checklist I I.

I suspect that in most situations, the answer to the first question will be that the two populations are disjoint.

Checklist III.

These questions concern whether the two populations inhabit the same geographic area. Presumable the interest is in similar climate, activity patterns, etc. Spatial characteristics convey a broader-in fact a different-meaning to me. It suggests how the population is distributed in space. Is it a high density area or a low density area? Are there clusters of housing separated by open space?

Responses to the Questions on Representativeness

Issue Paper on Empirical Distribution Functions and Non-Parametric Simulation

Introduction

Is stochastic variability really the right term here? Just to make sure I am interpreting this right, I take "variability" to mean that, for example, some people drink more tap water than others and thus have a greater exposure. The big difference between variability and scientific uncertainty or random error is that it is presumably possible to identify which individuals drink 2 liters/day and which drink 0.5 liters/day, or they can identify themselves. This is important because it provides a tool for intervention. For example, we can warn pregnant women to reduce their intake of fish rather than setting a standard requiring everyone to eat fewer fish. "Stochastic variability" seems to imply variability that is so randomized that we-nor the individuals involved-cannot determine who has a high exposure and who has a low exposure. In that sense, it is the same as a cancer dose-response function.

Why do we write-off the use of theoretically based distribution functions? Many environmental variables do seem to be distributed lognormally. It isn't just coincidence. I believe that we are often better off fitting our data to a lognormal than trying to develop an empirical distribution based on what is typically a rather small data set. I once got some good advice when I was a junior engineer trying to figure out how much water was flowing in a pipe. My boss told me, "We have a good theory explaining the flow of water in pipes, but our meters have a 5% error at best. If there is a difference between the theory and the data, assume the meters are wrong." My only problem with lognormals is how well they continue to map nature out in the extreme tails. Even there, however, how much confidence do we have in the 99th percentile of an empirically based distribution?

Part 1. Empirical Distribution Factors

Extended EDF

The EDF is extended by adding plausible lower and upper bounds, but the paper does not mention how one extends the linearized curve to reach those bounds. Presumable by using a curve-fitting routine of some kind.

In many cases, there is no clearly obvious point for the upper or lower bound. We know we do not have any one kg adult males, but how do we decide to stop at 15 kg and not 14? Expert judgment is used. Expert judgment may be all we have, but it is not a great justification, and it is important that we provide justification. I believe it is worthwhile to do a sensitivity analysis to find the difference between using quasi-arbitrary bounds and letting the curve run out to zero or infinity. It might also be worthwhile to check the difference with stricter, but perhaps more reasonable bounds, say a 40 kg adult male.

Samuel Morris

Mixed Empirical-Exponential Distribution

I think that mixing theoretical distributions with empirical distributions in some kind of composite sounds like a good idea.

Starting Points

The smaller the data set, the greater the rationale for using a standard distribution.

Responding to #5, people feel more comfortable with a theoretical distribution because it has a theoretical basis that supports interpolation between data points and extensions beyond the data, although I was always told never to do the latter. When plotting empirical data without a theory, one never knows if there is some big discontinuity between two completely innocent looking data points. The problem is that the theory behind the distribution is mathematical, not physical. To be comfortable with interpolating or extrapolating in either case, one must have a theory of the physical process involved.

P. Barry Ryan

Workshop on Selecting Input Distributions for Probabilistic Assessment

In the transmittal letter dated March 27, 1998, Beth O'Connor asked us as reviewers to provide "… not... comprehensive comments, but rather your initial reaction and feedback on the issues... ." Further, we have been asked to focus on the so-called "Representativeness" Issue Paper. My discussion focuses on that manuscript to start.

First Reactions

My first thoughts on this paper center on the need for an "audience" to be selected. Issue papers such as this one will lead, eventually, to guidance documents similar to those supplied as background reading. But what is the audience of this document? To a degree, the audience must be viewed as one and the same. This document will be referenced in a guidance document. Assuming this, a diligent worker looking for more information will seek out this manuscript. Hence it should be readable and accessible to practitioners of risk assessments and exposure assessment science. With this assumed audience in mind, I continue with my initial reaction to the Issue Paper.

The **Introduction** commences with a single sentence that concisely described the purpose of the document. This is a good start; the reader is entitled to know what is being discussed. Unfortunately, the next sentence is a parenthetical notation. Is this statement unimportant, less important, to be ignored, or what? The third sentence has a relative pronoun as the first word but the antecedent is unclear. To what does "This" refer? Exposure factors? Representativeness? Whatever it may be, it is both extremely brad and extremely important as the rest of the sentence tells us.

Before the above is dismissed as grammatical nitpicking consider the following. At this point, we are only three sentences into the document and I, considered to be an expert

reviewer, am uncertain as to what is being discussed. A gentle introduction to a difficult subject goes a long way toward keeping the reader “on line.” A little editing for style up front will make this document much more useful.

Let us continue. The next paragraph is a roadmap describing the way through the remainder of the document. These two paragraphs provide the **Introduction.** More is needed. Why is this important? When should it be applied? What has been done in the past? These are all reasonable questions to ask.

The next section begins the meat of the Issue Paper. **General Definitions/Notions of Representativeness** is a real mouthful of a title. The term “Notions” has the connotation of uncertain knowledge. Definitions are quite the opposite. Will we be treated to contradictory information in this section? Apparently the answer is ‘Yes” because, as pointed out the Issue Paper continues, a reference to Kruskal and Mosteller indicates that the term on which we are seeking guidance has no "... unambiguous definition..." Why is it necessary so early on in the discussion to confuse the issue in the mind of the reader by saying that no definition exists? Why would a reader of this document continue reading rather than throwing his or her hands up in despair?

The next paragraph (and accompanying table) adds further fuel to the fire. What is the purpose of this table? How does it contribute to the definitions or notions of representativeness? There is no discussion of the importance of the terms, how they might be used in assessing representativeness, nor the purpose of the table.

So, again, we have a section that needs significant editing. It is not clear to me that this section adds any insight into the notion (or definition) of representativeness. The elementary concept is not difficult. The attempt to be all-inclusive at the very beginning, however, is doomed to failure. It is difficult to tell someone what works by telling him or her all of the problems with the system first. It would be better to adopt a working definition, show how it can be applied to many situations, then list some problems with the working definition. This allows the reader to gain some understanding of the concepts, without having to grasp the entire subject a *priori.*

I have, until this point, spent a great deal of time discussing a very small part of the Issue paper. In particular, I may have spent more space on the discussion than the manuscript length to this point. However, the first page or two of any document sets the tone for the whole piece. The tone for this manuscript ranges from one of despair to one of disorganization. There is very little room in that continuum for gaining new insight. I urge a re-write of these early sections.

Moving on to the next section, **A General Framework for Making Inferences,** begins the “meat” of the manuscript. As a matter of style, I do not care for a series of parenthetical notations in sentences. I believe that it obscures the meaning of the prose. Shorter sentences fully describing each of the activities are better. This is a recurring style point throughout the document. I will not comment on it further.

Figure 1 represents a nice, concise “decision tree” approach to risk assessment data collection. The discussion is muddied somewhat by the introduction of the (undefined) concept of surrogate data. Reordering of sentences in the paragraph to bring the example closer to the first use of the word surrogate would clarify substantially. But we quickly go far afield from our discussion of representativeness. The manuscript needs